A Natural Language Understanding System Combining Syntactic and Semantic Techniques

Peter Haug, Spence Koehler, Lee Min Lau, Ping Wang, Roberto Rocha, and Stan Huff
Department of Medical Informatics, LDS Hospital, Primary Children's Medical Center, and the University of Utah, Salt Lake City, Utah

A large proportion of the medical record currently available in computerized medical information systems is in the form of free text reports. While the accessibility of this source of data is improved through inclusion in the computerized record, it remains unavailable for automated decision support, medical research, and management of medical delivery systems. Natural language understanding systems (NLUS) designed to encode free text reports represent one approach to making this information available for these uses. Below we describe an experimental NLUS designed to parse the reports of chest radiographs and store the clinical data extracted in a medical data base.

INTRODUCTION

The last decade has witnessed a growing awareness of the usefulness of computers in medical care. With this awareness has come increasing pressure to capture and store clinical data in computerized medical records systems. Several groups have reported significant accomplishments in developing a computerized medical record both in the inpatient and outpatient settings [1,2,3].

One of the central challenges in this process is that of capturing clinical data in a form that serves the needs of a variety of different information consumers. The first of these consumers are the physicians, nurses, and therapists that provide day to day care for the patient. Ease and speed of access are the principal goals of this group, but a characteristic of growing importance to these information users is the ability to drive medical decision support from the information in the data base.

A second group of consumers are medical researchers. These information users expect clinical data to be stored in a form that will support investigation into the science of medicine and health care delivery.

A third, and an increasingly important group of medical data users are the people who administer health care systems. They expect to use the information available in attempts to modify the ratio of benefits to costs in medical care delivery systems. To serve their needs clinical data must be collected in a way that supports quality assurance initiatives, health care planning, and computer administered protocols to help standardize the health care product.

These groups are joined by the federal government as well as other third party payers in a desire for more and better data with which to monitor health care in the United States.

The needs and goals of these information consumers are best served by data that is stored in a carefully encoded form defined in a controlled medical vocabulary (CMV). A variety of CMVs have been developed and various groups are promoting an effort to define the essential terminology for a medical vocabulary as well as the basic characteristics of clinical data storage on a national basis [4,5].

Unfortunately, a large proportion of the information that finds its way into the medical record consists of free text data. This includes highly relevant information in reports of the history and physical examination, accounts of x-ray examinations, pathology reports, the narrative descriptions of surgical interventions and other invasive procedures, and the condensed description of hospitalization contained in the discharge summary. To fill the needs of the groups mentioned above this data must be encoded secondarily.

Several groups have evaluated techniques for automatically encoding textual documents from the medical record. The Linguistic String Project has developed a series of tools for analyzing medical text [6,7]. Gabrielli has described a system for encoding discharge summaries for quality assurance [8].

X-ray reports appear to have a special appeal. Two groups have developed systems whose focus is the radiologists' report of the chest x-ray. Zingmond has applied a semantic encoding tool to these reports to recognize abnormalities that should receive follow-up [9] and Friedman

0195-4210/94/$5.00

has studied techniques for encoding interpretations found in these reports [10].

We have been using a semantic parser for five years to encode salient features from the reports of chest radiographs [11,12]. While the accuracy of this system is far from perfect, the results have been adequate to support a computerized expert system for screening nosocomial infections [13]. At present, we are actively involved in the development of an experimental natural language understanding system (NLUS) designed to answer a set of questions concerning the synergistic relationship between semantic and syntactic parsing techniques.

A NLUS whose goal is to read medical text and to extract and encode the clinical data embedded in this text has one basic requirement. This requirement is a model of the data representation into which the encoded data will be stored. A data model serves to provide both a target for the parsing process and to identify and circumscribe the set of concepts which will be managed by the NLUS system. We have chosen to use a controlled medical vocabulary and set of data structures known as the event definition model [14] as the target for a new medical parser. To this we have added a syntactic parser based on an augmented transition network grammar [15] and a semantic grammar managed as a Bayesian network [16]. These constituents are described below.

EVENT DEFINITIONS

The event definition data model consists of a dictionary whose purpose is to define not only the medical lexicon used in the data base, but also to specify salient structural components of the data base. Slots are defined and their relationships to each other and to objects called event definitions (ED) are cataloged. The extended dictionary in which this occurs is referred to as the master object index (MOI).

Specific medical facts are recorded using the event definitions themselves. These structured multi-slot objects provide the basic framework in which atomic concepts defined in the MOI are integrated into complex concepts adequate to represent instances of clinical data or events. Figure 1 shows an instantiated event definition, drawn from the realm of chest radiology. This example representing the medical event documented in the sentence, *"A hazy opacity is seen in the right upper lobe."*

The goal of the NLUS engine described below is to parse sentences like this one, to properly instantiate event definitions, and to store these event definitions in a general purpose medical data base. The process is controlled by a syntactic parser.

*Finding Event: *Localized Infiltrate*
- *State: *Present*
  - Presence Marker: *demonstrates*
- *Finding Unit: *Poorly-marginated opacity (infiltrate)*
  - Finding: *opacity*
  - Finding Modifier: *hazy*
  - *Severity: *null*
- *Anatomic Unit: *Right upper lobe*
  - *Link Unit: *Involving*
    - Anatomic Location Link: *in*
  - Anatomic Location: *lobe*
  - Sidedness Modifier: *right*
  - Superior/Inferior Modifier: *upper*
- *Change Unit: *null*

Figure 1: An instantiated event from a chest x-ray report. The slots indicated with a * are higher-level concepts found in a controlled medical vocabulary. The other slots are holders for words from the sentence. "Null" slots at the word level are not shown.

SYNTACTIC TECHNIQUE

The NLUS developed for this experiment is based on a set of augmented transition network (ATN) grammars [15] and a lexicon derived from the Specialist Lexicon developed at the National Library of Medicine [17]. This lexicon has been augmented with a group of multi-word phrases representing frequently seen combinations with standard meanings (i.e. "consistent with", "no significant"). A small list of synonyms is used to replace words that represent a combination of concepts with the specific concepts (i.e. "cardiomegally" = "enlargement of the heart").

The ATN grammars are used in a cascaded fashion. A first grammar is applied to constraint the syntactic identity of the individual words of a sentence. The syntactic classification of a word is constrained to a single category based on the syntactic categories of its neighboring elements. For example, the word "project" could be classified as a noun or as a verb. Given the two words "the project", the classification would be constrained to noun since "project" follows the article "the". Given the two elements "will project", the classification would be constrained

to verb since it follows a verb auxiliary.

The goal in this stage of syntactic classification is to find a single syntactic interpretation for each word which is mutually consistent with the categories of its neighbors and to determine which groups of words can be additionally categorized as higher level syntactic elements such as noun phrases. These groups of words will be referred to as constituents of the sentence. Note that for some sentences multiple syntactic interpretations may be possible.

Upon successfully recognizing a constituent, the NLUS collects the words comprising that constituent and bundles them as a single element. This element is classified using the appropriate higher-level syntactic category (noun-phrase, prepositional-phrase, etc.). Constituent grammars that use these phrase level syntactic assignments can then be applied until the sentence is completely categorized. The resulting structure has a one to one correspondence with a syntactic parse tree.

Ordering the application of constituent constraining ATNs is itself accomplished by an ATN with a constituent application grammar. This concept is similar to that of a cascaded ATN where the output of one ATN becomes the input of another.

A final step in the process of syntactic analysis is a transformational step aimed at producing a set of structures that match the needs of the semantic grammar. The principal goal here is to accurately associate those components of the sentence that, when combined, completely specify the clinical events represented in the sentence. The principal target of the transformations are the conjunctions found in these reports. The results are groups of syntactically categorized words and phrases divided into subsets (sentence fragments) likely to represent semantically meaningful utterances.

SEMANTIC APPROACH

The semantic knowledge used by this NLUS is stored and applied in the form of a Bayesian Network [16]. Figure 2 shows one of the experimental networks which we have used in testing. It is designed to represent the subset of information in a chest x-ray report used to indicate the abnormalities which have been seen on the film. Each of the nodes in the network represents a specific slot from the event definition. Leaf nodes provide place holders for individual words or phrases from the x-ray report. The intermediate and root nodes are associated with slots for higher level concepts defined in the MOI.

We use the Bayesian network representation in two ways. First, the structure of the network is used to indicate the relationship between the words from the sentence and the concepts associated with these words. For instance, the network in figure 2 indicates that an **Anatomic Unit** can be represented by words from the **Anatomic Location, Anatomic Location Modifier, Sidedness,** and **Inferior/ Superior** nodes plus a concept from the **Link Unit** node. Relationships among sets of concepts are also captured within the network. Each node has a restricted set of words or concepts which it can represent. This, combined with the network structure, allows the parser to restrict slot fillers to the limited set that can be used to express the concepts native to chest radiology.

Second, the probabilistic behavior of the network tends to further restrict word and concept assignments within the slots to those that are semantically meaningful. For example, the leaf nodes associated with the **Finding Unit** node could be successfully filled with words from the fragment, "*a hazy infiltrate*" and the words from the fragment "*of the right heart*" could fill the word nodes under **Anatomic Unit**, but the network would give a zero probability if both sets of leaf nodes were instantiated together. This would simply indicate that the finding "*a hazy infiltrate of the right heart*" should be considered semantically unacceptable. Similar behavior exists within the subtrees. The network would reject *"inferior cardiac enlargement"* because the combination of "inferior" with "cardiac" does not produce a meaningful probability for any concept in the **Anatomic Unit** node.

The semantics embedded in the Bayesian networks are invoked at the beginning of the parse of each sentence to set expectations for the syntactic parser and at the end of the syntactic analysis to test the set of slot instantiations which have been produced. Typically a group of possible instantiations are proposed by the syntactic processor and the best of these is chosen by the semantic grammar.

We are developing individual Bayesian networks to represent the semantics of each of the event types seen in chest radiology. In addition to the

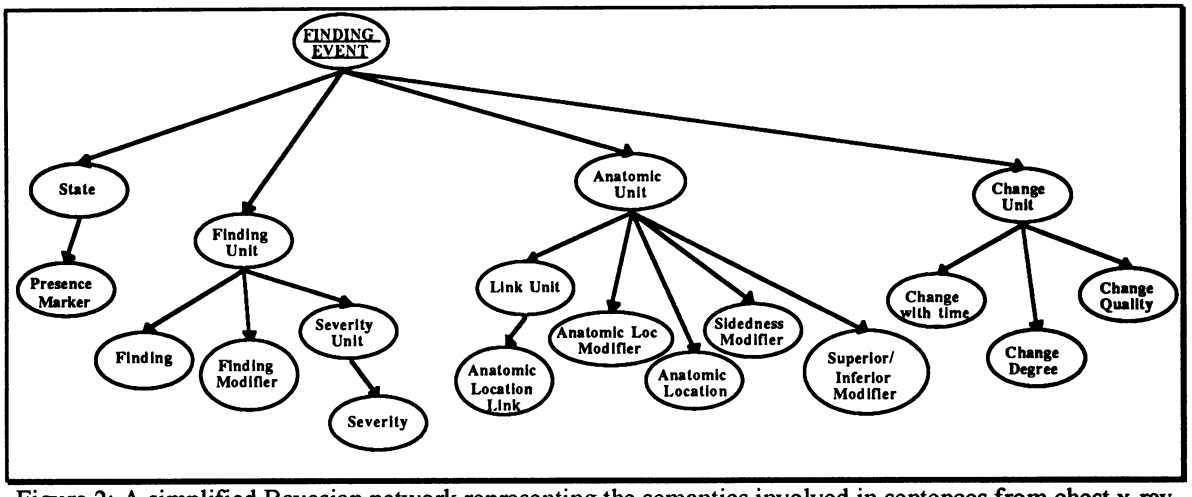

Figure 2: A simplified Bayesian network representing the semantics involved in sentences from chest x-ray reports describing findings.

networks for Findings, we have a network representing pathophysiologic interpretations (principally diseases) as well as a network for the various tubes and other hardware frequently described in radiologist examinations. The semantic and syntactic approaches which have been developed can clearly compliment each other even more effectively if proper ways of linking them can be developed. This is a subject which we are currently exploring.

We are developing this system through a series of small formative studies designed to focus our attention on the complexities of the syntactic-semantic relationship. To test progress we feed individual sentences into the parser and examine the ability of the system to recognize the key top level concept as well as the group of level two concepts that are associated with this key concept. In the example in figure 1 the key concept is the **Finding Event** and the second level concepts are the **State,** the **Finding Unit,** the **Anatomic Unit,** and the **Change Unit**.

RESULTS

In an analysis of ten chest x-ray reports collected sequentially from the HELP hospital information system, we found 50 sentences. Thirty-one of these sentences contained a total of 42 relevant key concepts. Nineteen expressed information outside of the realm of pathologic findings. The system correctly recognized the primary concept in 34 of the 42 cases (81%). Of 168 second level concepts, the system recognized 133 (79%).

The system's greatest current failing is a tendency to find concepts in sentences where they are not present. In this set of reports it generated 18 erroneous conceptual groupings. However, all but one were in sentences that dealt with the presence of various tubes, elements of patient history, and details of the radiologic procedure itself. We will soon begin testing with multiple Bayesian networks, each covering a different set of these concepts. As we begin using frameworks capable of encoding these other forms of conceptual abstraction, we expect a decrease in the frequency of these concept assignment errors.

DISCUSSION

The value of clinical data, accurately encoded using a CMV, has been demonstrated multiple times. In order to encode the medical data currently collected in a text-based form, system designers can attempt to replace natural language centered tools with structured interfaces designed to capture coded clinical information directly. The success of these interfaces has been limited in the past, particularly when the physician is the primary data source. The alternative is to supply tools capable of taking unstructured textual information, extracting salient facts, and encoding them.

The accuracy seen in the experiments described above does not yet match that of the semantic parser which we have been using. This tool has demonstrated a true positive rate of 87%-90% for chest x-ray findings similar to those in this study[12]. Unfortunately, this application has proven difficulty to maintain and cannot currently support the types of semantic and syntactic extensions which we wish to test. Because it was designed principally to explore a set of semantic theories, re configuring it to include syntactic knowledge and to properly

integrate the syntax and semantics would be difficult.

The experimental system described above is fully configurable. Both its syntactic and its semantic knowledge are stored separate from the program and can be altered to match the needs of other types of medical free text. We intend to use this feature to provide access to coded data from other natural language documents in the medical record.

Our experience with the nosocomial infection monitor has led us to design several new applications that depend on encoded free text. The first of these is a system for computer-assisted antibiotic ordering. It uses information extracted from the chest x-ray report to determine the character and duration of pulmonary infections. It is currently being tested in our intensive care units.

The second new application is a system for determining the problem that brought each patient to the hospital. This information is entered as free text at the time of admission and is later encoded by the Medical Records Department. In order to expedite medical decision support we are planning to parse and encode this data at the time of admission.

The ultimate goal of this development effort is to allow computing systems full access to text-based medical information. Creation of robust NLUSs will bring us a step closer to medical information systems that can act as full participants in the process of health care delivery.

* This publication was supported in part by grant number 5 R01 LM05323 from the National Library of Medicine.

References

1. McDonald CJ, Tierney WM, Overhage JM, Martin DK, Wilson GA. The regenstrief medical record system: 20 years of experience in hospitals, clinics, and neighborhood health centers. MD Computing (1992); 9:206-217.
2. Kuperman GJ, Gardner RM, Pryor TA. *HELP: A Dynamic Hospital Information System.* 1991 Springer-Verlag, New York.
3. Clayton PD, Anderson RK, Hill C, McCormack M. An initial assessment of the integrated academic information system (IAIMS) at Columbia Presbyterian Medical Center. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. (1991): 109-113.
4. Committee on Improveing the Patient Record, Institute of Medicine. *The Computer-Based Patient Record.* Ed. Dick RS, Steen EB. National Academic Press, Washington, D.C., 1991.
5. Board of Directors of the American Medical Informatics Association. Standards for Medical Identifiers, Codes and Messages Needed to Create an Efficient Computer-Stored Medical Record. American Medical Informatics Association JAMIA 1994; 1:1-7.
6. Sager N, Friedman C, Lyman MS. *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley, Menlo Park, CA (1987).
7. Sager N, Margaret L, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. JAMIA (1994); 1:142-160.
8. Gabrielli ER. Computer assisted assessment of patient care in the hospital. J Med Syst(1988); 12:135.
9. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comp Biomed Res 1993; 26:467-481.
10. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. JAMIA 1994; 1:161-174.
11. Ranum DL, Haug PJ. Knowledge based understanding of radiology text. In: Greenes RA, ed. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. (1988): 141-145.
12. Haug PJ, Ranum DL, and Frederick PR. Computerized Extraction of Coded Findings from Free-Text Radiology Reports. Radiology (1990); 174:543-548.
13. Evans RS, Gardner RM, Bush AH, Burke JP, Jacobsen JA, Larsen RA, Meier FA, Warner HR. Development of a computerized infectious disease monitor (CIDM). Comput Biomed Res (1985) 18:103-113.
14. Huff S. Medical Data Dictionary for Decision Support Applications. Proceedings: Eleventh Annual Symposium on Computer Applications in Medical Care (1987); pp 310-317.
15. Woods WA. Transition Network Grammars for Natural Language Analysis. Communications of the ACM (1970); 13:591-606.
16. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, San Mateo, CA, 1988.
17. McCray A. Personal communication.